

CENSORED DATA: WHAT'S THE AVERAGE OF UNKNOWN VALUES?

By James R. Chastain, Jr., PhD, PE, MPH



Monitoring plans are an essential part of effective environmental regulation and management. The essence of any monitoring plan is to identify changes in key variables and use that information to assess the condition of the system of interest. Monitoring programs generally are divided into three categories, depending on their purpose. *Baseline Studies* seek to document the current state of the environment which provides a basis for quantifying changes in the future. *Targeted Studies* are used to assess the impact of planned events or quantify the effects of past events (such as a chemical spill or natural disaster). Finally, *Compliance Monitoring* is intended to detect trends in variables of concern and to document that the source is functioning within applicable guidelines or regulations. While the objectives of these programs vary, they all depend on well-defined plans to properly sample, analyze and interpret the data.

Unfortunately, data from environmental monitoring plans is notoriously “messy” or, as statisticians say, “not well-behaved”. There are a number of reasons for this. First, environmental data is typically both temporally and spatially variable. Further, there may be numerous parameters to track and define. Also, it is not unusual for human or equipment error to result in lost or spurious data. When correlations between variables are important, these and other effects complicate the interpretation significantly. One of the primary tasks that one faces when evaluating a data set is to develop summary statistics which includes a measure of central tendency and variance. This can be difficult if the monitoring data includes a significant proportion of censored data.

Censored data most commonly occurs when the true value of the sample lies below the detection limit of an analytical test. While there are some additional subtleties [ex. Limits of Quantification, Method Detection Limits (MDL), Practical Quantification Limits, etc.] basically this means that the laboratory report comes back with a sample concentration of “not detected” or “< MDL”. In one sense, the result is probably good news because it most likely means that the system is in compliance with some stipulated standard. However, when trying to compute statistics or trends, the actual value is indeterminate. Many times it is also of interest to examine correlations between parameters. When censored data is encountered, especially in conjunction with multiple variables, the analysis becomes more complicated and prone to bias, if not error. Given the frequency that this issue occurs, what is the appropriate way to handle censored data?

The approaches can get quite complicated. For the purpose of this article, we will consider techniques that are typically used when estimating the mean and standard deviation, although they can be extended to computations of other statistics also.

To reiterate, the objective of this exercise is to determine a number that lies between zero and the MDL that can be used to compute the mean and standard deviation of a list of values. A summary of the more common approaches is as follows:

Simple substitution: this is probably the most common resolution used to address the problem. Typically, a fixed value is chosen from the interval between zero and the MDL and then substituted in any place where a censored value occurs. If zero is used, it results in a negative bias because that is the lowest value possible. On the other hand, if MDL is chosen, the analysis will have a positive bias since that is the maximum value the point(s) can assume and still be a non-detect. Many analysts, therefore, use half of the MDL to hedge the bias of selecting a more extreme value.

Direct Maximum Likelihood Estimator (MLE) methods: These methods, initially developed by Cohen, assume that the unknown values follow a given distribution and therefore can be estimated by maximum likelihood techniques. MLE techniques are fairly well documented in statistics literature. Cohen and those that followed have developed a series of tables or computer algorithms to compute the appropriate values.

Regression on Rank Order Statistics: In this non-parametric procedure, all values are rank-ordered with the “<MDL” values listed as smallest and the percentile computed for each value. The value is then plotted against its probability (z-score). Only the non-censored values can be plotted, of course but, if the resulting line is approximately straight, the censored values can be estimated by interpolation. Log transformation is also possible if the probability plot is not straight. This process is straightforward and can be adapted to a computer spreadsheet. The regression features can then replace graphical interpolation.

Robust Parametric Method: This method is similar in concept to the procedure above except that this method is a parametric procedure. That is, this method assumes that the data follows a stipulated distribution (usually a normal or log-normal form). Again, a probability plot is constructed using the uncensored data. The censored values are then replaced by extrapolated values from the fitted regression line.

Method of Proportions: Taking a page from categorical data analysis, another technique for examining the data set is to classify the data set only as “values above MDL” and “values below MDL”. This procedure is generally used when the majority (> 50%) of given data is below the detection limit but at least 10% of the observations are quantified. Because of the extreme level of censoring in this case, typically a percentile slightly greater than the proportion of non-detects is used for the confidence interval rather than the mean.

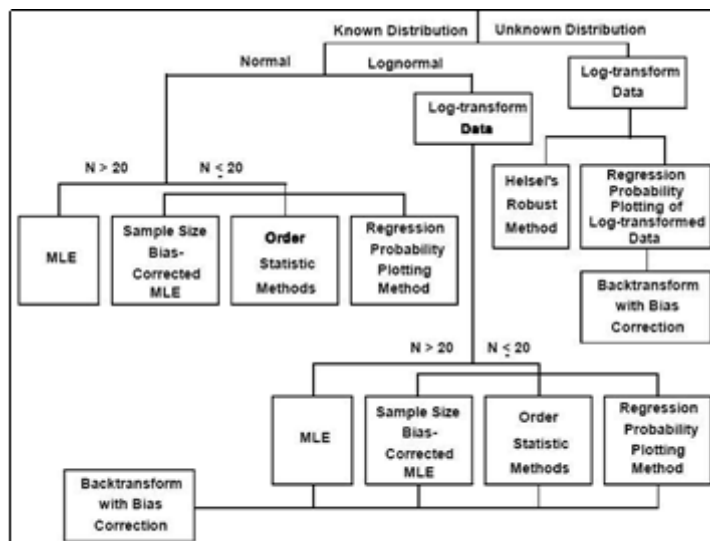


Figure 1

These, along with trimmed means, Winsorized means, Aitchison’s Method and other tests, provide a wide array of methods to deal with this situation. This all begs the question, which test should be used?

The answer, in true statistician fashion, comes back...that depends. The method used depends significantly on (1) the total number of observations in the monitoring cycle, (2) the proportion of those tests that are below detection limits, (3) whether the distribution of the data is known or unknown, and (4) what statistics are being required (ex. means, confidence intervals, trends, etc.)

The good news is that computer programs are available to manage the tedious computations. However, it is necessary to have an understanding of the analysis options and their strengths and weaknesses in order to select the appropriate alternative. Because most standard statistical programs seldom have many of these analyses as part of their standard menu, a public domain computer program named UNCENSOR (Newman *et al*, 1995) is

quite useful for those who know how to use it. The program can be downloaded from the Virginia Institute of Marine Science (College of William & Mary) website: www.vims.edu/env/research/software/vims_software.html

The User’s Manual has a flow diagram that helps guide the method selection process which is shown as Figure 1.

Practically speaking, it may not be necessary to go through an elaborate analysis for non-critical monitoring events. The Environmental Protection Agency (EPA) in some of their technical guidance has proposed a more relaxed procedure for routine analysis (USEPA, 2006). This approach suggests that the analysis method be selected on the basis of the percentage of non-detects in the data group.

The major purpose of this article is to be mindful of the fact that the handling of data below the detection levels requires purposeful evaluation. For those using automated spreadsheets to manage their data, this topic certainly warrants a review of the computational procedure to confirm that this issue is being properly handled. Even though censored data complicates an analysis, the information should never be thrown out.

Approximate Percentage of Non-Detects	Statistical Analysis Method
<15%	Replace non-detects with 0, MDL/2, or MDL; Cohen's method
15% - 50%	Cohen's method, Winsorized mean/std. deviation
> 50% - 90%	Proportions Test

References:

Manly, B.F.J. (2001). *Statistics for Environmental Science and Management*. Chapman & Hall/CRC. Boca Raton, FL.
 Newman, M.C., Greene, K.D., Dixon, P.M. (1995). *UNCENSOR*© v.4.0. Savannah River Ecology Laboratory. Aiken, SC.
 USEPA (2006). *Data Quality Assessment: Statistical Methods for Practitioners*. EPA/240/B-06/003. Washington, DC.
 Dr. Jim Chastain is the President of Chastain-Skillman, Inc. He has a Bachelor of Science in Civil Engineering (honors) and Master of Engineering from the University of Florida. He also has a Master of Public Health and Ph.D. from the University of South Florida. He is a registered Professional Engineer with over 30 years of experience and is a Diplomate of the American Academy of Environmental Engineers. He can be reached at (863) 646-1402 or jrchastain@chastainskillman.com.